Plenarsitzungen

Computerbiologie – die Welt der großen Zahlen und Moleküle*

Dietmar Schomburg

Institut für Biochemie, Biotechnologie und Bioinformatik der TU Braunschweig
Spielmannstraße 7, D-38106 Braunschweig

Biowissenschaften – von der deskriptiven zur quantitativen/synthetischen Wissenschaft

Historisch war die Biologie mit Ausnahme der Biochemie länger als ihre Nachbardisziplinen eine deskriptive, systematisierende Wissenschaft. Im Bereich der Biochemie entwickelten sich in der zweiten Hälfte des 20. Jahrhunderts hocheffiziente analytische Methoden, die in den letzten Jahren zu einer Explosion unseres biologischen Wissens führten. Erwähnt sei hier nur 1949 die Bestimmung der ersten Proteinsequenz, des Insulin, die Bestimmung der ersten Protein-3D-Struktur im Jahr 1958 und 1972 die Bestimmung der ersten Gensequenz. Es folgten die systematische Sammlung von Gen- und Proteinsequenzen erst in Buchform, dann in Form von Datenbanken, die Entwicklung von Rechner-basierten Sequenzvergleich und Sequenzalignment Programmen. In den letzten Jahren erleben wir eine extrem schnelle Weiterentwicklung von analytischen Methoden, die uns die Gesamtinformation über die jeweils aktuelle Zusammensetzung einer Zelle in Form von Genom, Transcriptom, Proteom, Metabolom geben.

Möglich wurde diese Entwicklung nur durch die parallel laufende schnelle Entwicklung der Computerbiologie (Bioinformatik, Systembiologie), die wiederum von der Entwicklung der Computer-Hardware abhängig war. Möglich wurde die komplette Genom- und Genfunktionsanalyse heute von mehr als 2000 Organismen, daneben die Simulation ganzer Zellen, das Design von Medikamenten und Proteinen mit neuen Eigenschaften.

Die Computerbiologie – oder Bioinformatik – befasst sich heute mit ganz verschiedenen Anwendungsfelder,

– Biologische Information, ihre Gewinnung und Integration
– Biologische Strukturen
  (Metabolite, Makromoleküle, Sequenzen, 3D-Strukturen, Zellen, Organe)

---

* Der Vortrag wurde am 19.01.2013 vor der Plenarversammlung der Braunschweigischen Wissenschaftlichen gesellschaft gehalten.

– Biologische Funktionen, vor allem molekularer Netze/Zellen (Regulatorische Netze, metabolische Netze, Transport)

Biowissenschaften zwischen Molekül und Organismus

Seit dem Ende des 19. Jahrhunderts sind die Biowissenschaften insgesamt gespalten zwischen den makroskopischen Beobachtungen, der Beobachtung des sogenannten Phänotyp, wie z.B. ein Organismus überlebt bei 90°C, eine Pflanze braucht Schatten, ein Mensch hat Fieber oder eine biotechnologische Fermentation ergibt nicht genug Produkt und den molekularen Beobachtungen, dem sogenannten Genotyp, wie eine genetische Mutation liegt vor, ein Protein wird in höheren Mengen produziert, eine bestimmte Enzymaktivität steigt, es werden Wechselwirkungen eines Proteins mit anderen Molekülen beobachtet.

Hier deutet sich konzeptionell eine Überwindung dieser Spaltung an, die mit Hilfe der Systembiologie erreichte „Genotyp – Phänotyp Korrelation", die große Herausforderung der Lebenswissenschaften in den nächsten Jahren.

Vieles haben wir das in den letzten Jahren schon erreicht, bei den genetisch monokausalen Phänomenen, beispielsweise manche genetische Krankheiten wie Phenylketonurie, die heute oft schon verstanden und qualitativ vorhersagbar sind.

Die meisten biologischen Phänomene stellen aber multikausale und kooperative Phänomene, dar, die durch qualitative Betrachtungen nicht verständlich werden, wie z.B. die komplexen Krankheiten (Herz/Kreislauf, Krebs, Infektionen, Autoimmun-Krankheiten etc.). Hier liegen Reaktion von molekularen Netzen vor, nicht von einer oder weniger Komponenten. In dieser überwiegenden Mehrzahl der Fälle sind Modellierung und Simulation notwendig

Für diese Simulation braucht es aber eine Reihe von Voraussetzungen.

Wir brauchen genaue Kenntnis :

– Aller zellularen Komponenten
– Ihrer statischen und dynamischen Wechselwirkungen und Umwandlungen
– Der Netzwerktopologie
– Der Regulation
– Der Grenzen des Systems und der Transportwege
– Die mathematische Beschreibung und Software
– Kenntnis der physikochemischen Gesetze und Parameter sowie des jeweiligen Energiebedarfs (Kinetik und Thermodynamik).

Biologische Systeme sind groß

Allein die Anzahl der molekularen „Player“ in der Zelle zeigt, dass ein Erfassen oder gar ein Versständnis und eine Vorhersage der Vorgänge ohne Computer völlig unmöglich ist.

Die Zelle: molekulare Bestandteile der bakteriellen Zelle

	Gewichtsanteil	Zahl der Komponenten
Wasser	70%	1
DNA	1%	1
RNA	6%	>3000
Proteine	15%	3000
Lipide	2%	20
Polysaccharide	3%	5
Metabolite	2%	500–1000
Ionen	1%	20

Genome – Konstruktionszeichnungen & Produktionsanweisungen für Proteine

Die Mikroorganismen mit dem kleinsten Genom tragen ca. 600 000 Buchstaben („Basenpaare“), die für ca. 500 Gene kodieren, dies entspricht d.h. ca. 300 Seiten Text in einem Buch. Die normalen und häufigen Bakterien schon ca. 3–20 Millionen Buchstaben, die für 3000–20 000 Gene bzw. Proteine kodieren, d.h. ca. 3–20 Bücher mit je 500 Seiten Text, der Mensch hat ca. 3 Milliarden Buchstaben und ca. 25 000 Gene, d.h. ca. 3000 Bücher mit je 500 Seiten Text.

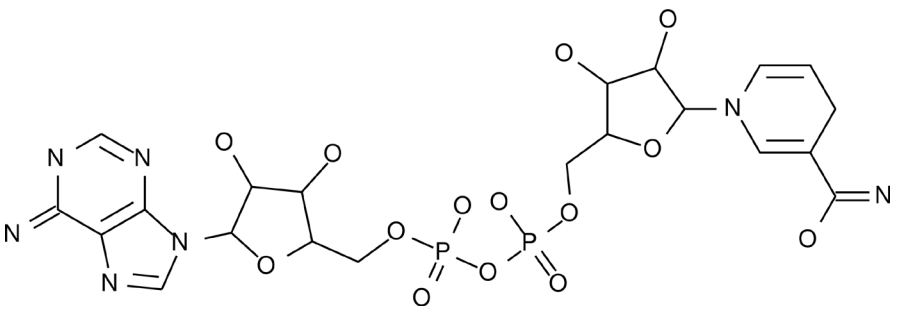
Manche anderen Organismen benötigen aber sehr viel mehr genetisches Material als der Mensch, so die Tulpe mit ca. 120 Milliarden Buchstaben, dies entspricht ca. 4 km Bücher nebeneinandergestellt oder der südamerikanische Lungenfisch mit ca. 784 Milliarden Buchstaben, d.h. ca. 25 km Bücher. Aber auch winzige Einzeller, z.B. die Amöbe *Amoeba dubia* mit ca. 670 Milliarden Buchstaben, ca. 22 km Bücher, schlägt den Menschen bei weitem.

DNA, Kohlenhydrate, Proteine und Metabolite – Funktionen dank definierter chemischer und Raumstrukturen

Während die Genome in ihrer Funktion eindimensional kodiert sind, definiert sich biologische Funktion der meisten Moleküle durch ihre Raumstruktur und die elektrostatischen Eigenschaften ihrer Oberfläche. Im Computer übersetzt in

entsprechende Kodierungen, wie z.B. der InChi-Code, buchstabenbasierte Gen- oder Proteinsequenz und die Koordiniertenbaierte Protein 3D-Struktur.

Als Beispiel sei hier NADH als chemische Strukturformel und in der Computer-lesbaren Form als InChi (International Chemical Identifier) dargestellt.



InChI=1S/C21H29N7O14P2/c22-17-12-19(25-7-24-17)28(8-26-12)21-16(32)14(30)11(41-21)6-39-44(36,37)42-43(34,35)38-5-10-13(29)15(31)20(40-10)27-3-1-2-9(4-27)18(23)33/h1,3-4,7-8,10-11,13-16,20-21,29-32H,2,5-6H2,(H2,23,33)(H,34,35)(H,36,37)(H2,22,24,25)/t10-,11-,13-,14-,15-,16-,20-,21-/m1/s1

Biologische Systeme sind dynamisch

Die Zelle reagiert zu jedem Zeitpunkt auf äußere Signale, entweder aus der Umgebung oder von anderen Zellen kommend. Dabei reagiert der Stoffwechsel am schnellsten auf die Umwelt. Daneben steht der Zelle die Möglichkeit zur Verfügung, über Regulations-Netzwerke ihr Reservoir an Proteinen optimal an die veränderte Situation anzupassen. Die kostet aber deutlich mehr Zeit und vor allem Energie, das die Zelle in Form der Hydrolyse von ATP Molekülen liefern muss. So kostet z.B. die Produktion des genetischen Materials 72 Millionen Moleküle ATP, eines Proteins 1.500, eines Polysaccharids oder eines mRNA Moleküls je 2.000 ATP Moleküle.

Molekulare Diffusionsgeschwindigkeiten in der Zelle sind deutlich herabgesetzt, so von größeren Proteinen um ca. den Faktor 10. Neben der Funktion ist auch die Lebensdauer von Molekülen in der Zelle während der Evolution optimal angepasst, so werden manche Proteine schon nach 10 Minuten, andere erst nach ca. einer Woche „recyclet“.

Die experimentellen Methoden

Diese Anpassungen und schnellen Umwandlungen in der Zelle haben wir bis vor kurzem nur am Einzelfall und in Ansätzen untersuchen können, aber die modernen

experimentellen Möglichkeiten, z.B. die Massenspektrometrie in Proteom- oder Metabolomanalysen und die modernen Sequenzierautomaten erlauben uns jetzt die Analyse, und mit der Analyse kommt das Verständnis.

Die schnellste technologische Entwicklung im Vergleich aller weltweit zur Zeit entwickelten Techniken überhaupt erlebt zur Zeit sicherlich die Genom-Sequenziertechnik. Während die Bestimmung des ersten menschlichen Genoms im Jahr 2003 ca. 2,7 Milliarden US $ gekostet hat, kostet die Sequenzierung des Genoms eines Menschen heute nur noch ca. 1000 $, d.h. millionenfach weniger.

Datenbanken, in der Biologie geht nichts ohne sie

Die erwähnte technische Explosion zieht natürlich eine Explosion des in öffentlichen Biodatenbanken abgelegten Wissens nach sich. Hierzu gehören sogenannte Primärdatenbanken, in der experimentell bestimmte Information wie Sequenzen und Strukturen abgelegt werden und komplexere Datenbanken, die Biomoleküle oder Organismen beschreiben.

BRENDA – umfangreiche Datenquelle für die Biowissenschaften

Als Beispiel einer seit 25 Jahren entwickelten biologischen Datenbank sei hier das Enzym-Informationssystem BRENDA (Braunschweiger Enzymdatenbank) erwähnt. Sie enthält experimentelle Daten von mehr als 1,4 Millionen Enzymen, von denen 67 000 experimentell charakterisiert sind. Insgesamt wurden aus manuell ausgewerteten 118 000 Journal-Artikeln ca. 3,4 Millionen Daten extrahiert und strukturiert zugänglich gemacht. Die enorme Bedeutung der Datenbank für die internationale Forschergemende wird durch die monatlichen 5–7 Millionen Zugriffe von ca. 200 000 Benutzern weltweit deutlich. Die manuell aufgearbeiteten Daten werden durch Computer-basierte Textinterpretationen, sogenanntes Text-Mining, ergänzt.

Systembiologie – vorurteilsfreie Analytik & Modellierung

Im Rahmen der Systembiologie werden vorurteilsfreie Messmethoden mit rechnerbasierten Simulationsmethoden verknüpft und liefern so ein Verständnis des Verhaltens von Organismen, z.B. was erlaubt es einigen Organismen, uns zu infizieren oder bei 100°C zu leben, die Vorhersage der Reaktion des Organismus auf Änderungen in seiner Umgebung (z.B. bei Infektion oder Medikamentengabe), die Vorhersage des Verhaltens von Mutanten. Diese systembiologischen Methoden werden zur Zeit vor allem in der Grundlagenforschung, der biotechnologischen Produktoptimierung und der Medikamentenentwicklung eingesetzt.

Die verwendeten Modellierungsansätze lassen sich in die folgenden Kategorien einteilen:

– Strukturanalyse des metabolischen Netzes
  - Identifizierung von wichtigen "Umschlagplätzen" – Ansatzpunkte für Medikamente oder für biotechnologische Produktbildung.

– Stöchiometrische Modellierung
  - Berechnung der metabolischen Kapazitäten und stabiler metabolischer Zustände in Zellen und Organen
  - Vorhersage von Stoffflüssen in einem gegebenen Umfeld und einem bestimmten Protein-Repertoir
  - Analyse der Lebensfähigkeit und Widerstandsfähigkeit

– Kinetische Modellierung
  - Zeitabhängige Reaktionen der Zelle auf äußere Veränderungen

Die Modellerstellung ist momentan noch sehr zeitaufwendig und arbeitsintensiv. Man geht dabei im Normalfall von einem sequenzierten Genom aus, sagt die biologischen Funktionen der Gene voraus, addiert bei einem metabolischen Netz die chemischen Reaktionen der Enzyme, Transportvorgänge und Energiebilanzen und testet dann die aus dem Modell berechneten Vorhersagen und Hypothesen des Modells am Experiment. Jedes dieser Modelle durchläuft dann eine Reihe von experimentell definierten Optimierungszyklen.

Während die kinetische Modellierung für absehbare Zeit noch an dem Fehlen von Zehntausenden kinetischer Enzymkonstanten nur für sehr kleine Systeme denkbar ist, ist die stöchiometrische Modellierung heute „state of the art".

$$\frac{dc_1}{dt} = \sum_i S_{i1} \cdot v_i - b_1 = 0$$

$$\frac{dc_2}{dt} = \sum_i S_{i2} \cdot v_i - b_2 = 0$$

$$\cdots$$

$$\frac{dc_N}{dt} = \sum_i S_{iN} \cdot v_i - b_N = 0$$

Sie beruht auf der Aufstellung von partiellen Differentialgleichungen für jede Komponente der Zelle, Wenn es wie in vielen Fällen um die Modellierung stabiler Zustände geht, vereinfacht sich das System, wie oben gezeigt, zu einem – unterbestimmten – linearen Gleichungssystem. Dieses kann unter Zuhilfenahme von Zusatzannahmen, wie z.B. dass der Organismus während der Evolution in Richtung auf effiziente Nutzung der Nahrungsquellen optimiert wurde, gelöst werden.

Als Resultat erhält man z.B. eine Vorhersage der Stoffflüsse in der Zelle, wie in der folgenden Abbildung für *Sulfolobus solfataricus*, ein thermostabiles Archaeon, und Phenolabbau gezeigt.